# Multinomial and Dirichlet-multinomial modeling of categorical time series

Erik Thorsén*

June 9, 2014

### Abstract

This Bachelor´s thesis considers two categorical time-series regression models, the multinomial logit regression model and the Dirichlet-multinomial regression model. The interesting differences between these two regression models lies in their respective variance-covariance structure where the Dirichlet-multinomial is more flexible. Aim was to present a mathematical description of the two models and compare them for a specific dataset.

The two models where fitted, under the assumption of no autocorrelation, for a case analysis on proportions of age-categories of reported rotavirus cases from Brandenburg, Germany. The proportions of age-categories where of interest because one would like to see if there was a age-shift from young children and infants (00-04) to elderly (70+), after the introduction of a vaccination programme for infants and young children in 2009 (Koch and Wiese-Posselt, 2011). Data show strong seasonal variation. A graphical presentation of the two categorical regression models show that they fit mean values almost equal. However, large differences are shown in the goodness-of-fit measures AIC and BIC and in predictive intervals, constructed by sampling. On average, a predictive interval for elderly (70+) using the Dirichlet-multinomial time-series regression contained approximately 90% of the observations while as the multinomial logit regression contained approximately 50-60% of the observations. These results display that when data shows tendencies of over-dispersion, the Dirichlet-multinomial regression model is more adequate to model proportions with than the multinomial logit regression.

*Bachelor thesis at department of Mathematical Statistics, Stockholm University, SE-106 91, Sweden. Supervisor: Michael Höhle. Email: Ethorsn@gmail.com

# Contents

# Chapter 1

# Introduction

Regression models for categorical time series have been introduced as a alternative to Markov chain modeling, integer autoregressive processes and discrete ARMA processes for analyzing discrete time series data (Fokianos and Kedem, 2003). Regression models for categorical time-series are growing in popularity in fields such as biomedical surveillance, business monitoring and industrial quality control. The advantage of these lies in the possibility to use theory for generalized linear models (GLM) and that no assumption of Markov property or stationarity needs to be made (Fokianos and Kedem, 2003). Time series assumes dependence between past observations and the present, i.e. autocorrelation. Due to the setting of this thesis there will be no analysis of this fact. Data will be used as if it consisted of independent observations.

This thesis is based on a study of two categorical time series regression models: the multinomial logit- and the Dirichlet-multinomial regression. The study comprises of a theoretical treatment and a case study from infectious disease epidemiology. The theoretical treatment includes the multinomial distribution, the Dirichlet-multinomial distribution and the corresponding regression framework which is to be used for model fitting. Their respective variance-covariance structure is of great interest. The multinomial distribution has a variance-covariance structure which is solely defined by the expected values. When data shows tendencies of over dispersion, this variance-covariance structure may not be adequate. It is not flexible enough. The Dirichlet-multinomial distribution allows for a more flexible variance-covariance structure. The aim of this thesis is to show that the Dirichlet-multinomial regression model has practical use when nominal data shows tendencies of over dispersion.

The outline of this thesis is as follows. First, a descriptive analysis of data from the federal state of Brandenburg, Germany, containing monthly reported rotavirus cases. Rotavirus is a virus of the stomach which is the leading cause of infectious diarrhea among infants and young children. It is estimated that each year a total of 111 million rotavirus episodes occurs which requires home care (Parashar et al., 2003). The virus is seldom fatal in developed countries though for older age-categories, where dehydration can be fatal, the virus can have unexpected consequences. Data was collected over 12 years, 2002-2013. In 2009 the state of Brandenburg introduced a recommendation of a vaccine against the rotavirus for young children. The vaccine programme should imply a decrease in the total amount of reported rotavirus cases. All reported cases have been divided in different age-categories, ranging from newly borns to 70+. Age-categories is an ordinal scale. In this thesis age-categories will be used as though it was nominal in order to analyze data with the multinomial logit- and Dirichlet-multinomial regression model. Proportions will be of interest as we want to see which age-group is affected worst by the recommendation of the vaccine. In the second and third chapter we will discuss the multinomial-, Dirichlet-multinomial distribution and their respective regression model. The derivation of the Dirichlet-multinomial will be presented. All calculations are shown in Appendix or otherwise stated. We fit the regression models in chapter four, using **R**, a free software environment for statistical computing and graphics (R Core Team, 2014), with the **MGLM** package (Zhang and Zhou, 2014). Last, a chapter for discussion of the results and what may be done in future studies is presented.

To ensure reproducibility of the analysis, keep **R** code and description closely together we used knitr (Xie, 2013), for presenting **R** code and text together.

## 1.1 Descriptive analysis of Rotavirus data in Brandenburg, Germany, 2002-2013

The data used in this study were acquired at the webpage `https://www3.rki.de/SurvStat/`. Data contains monthly reported rotavirus cases in the federal state of Brandenburg, Germany. A total of 41627 cases where reported in 2002-2013. The data is grouped in 15 age categories, presented in Figure 1.4. With 15 age categories and 144 months of observation, i.e. 12 years, a total of 41627 case reports are scattered over 2160 cells.

In 2009, the state of Brandenburg introduced a recommendation of a vaccine against the rotavirus for infants and young children, in accordance with WHO recommendations. The vaccine had been on the market since 2006 (Koch and Wiese-Posselt, 2011). A decrease is seen in the total count of reported Rotavirus cases after 2009. Figure 1.1 shows absolute value plotted over time.

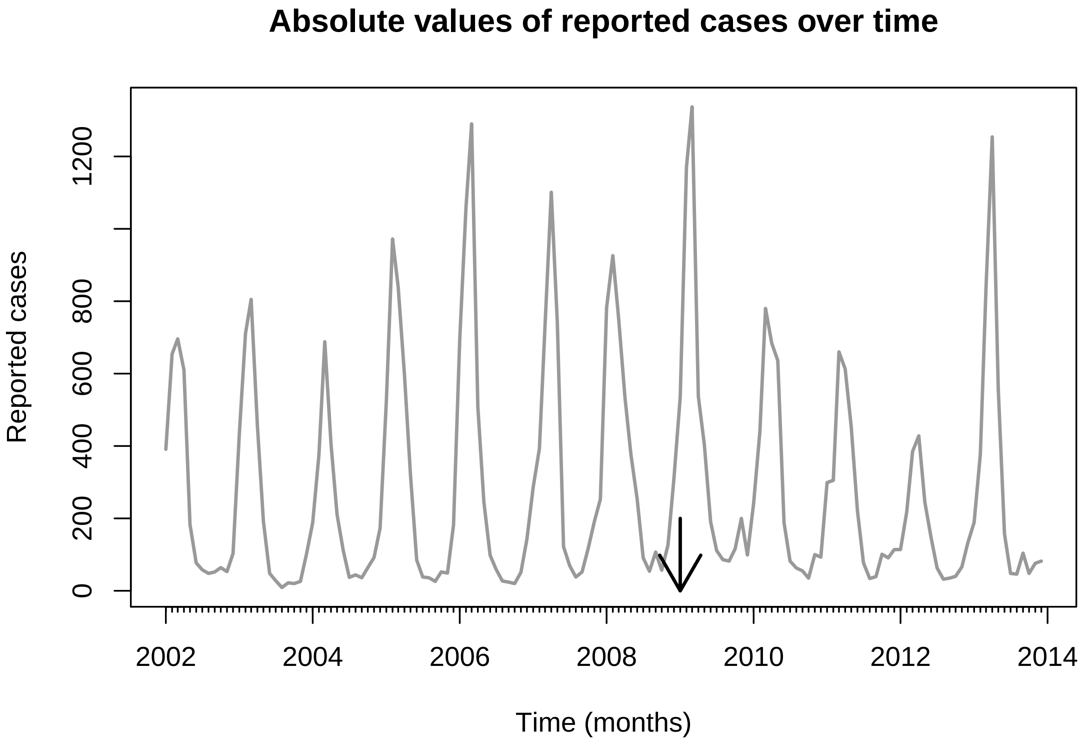

Figure 1.1: Number of reported cases over time, vertical arrow indicates the recommendation of the vaccine by the state of Brandenburg

The vertical arrow in Figure 1.1 indicates the introduction of the vaccine. Reported cases are seen to decrease after the introduction. However, an increase is shown in 2013. One also observes a large periodicity in data. In order to investigate the periodicity in our data we look upon the different years and how the cases are reported over the months. Note that a calendar year may be a insuffiecent period to properly show the periodicity. Peaks may occur at the start of each year. If so, maximas would be in the beginning or in the end of the calendar year. As a consequence, the graphical presentation would miss to display important information. Hence, a period of September to August is chosen.

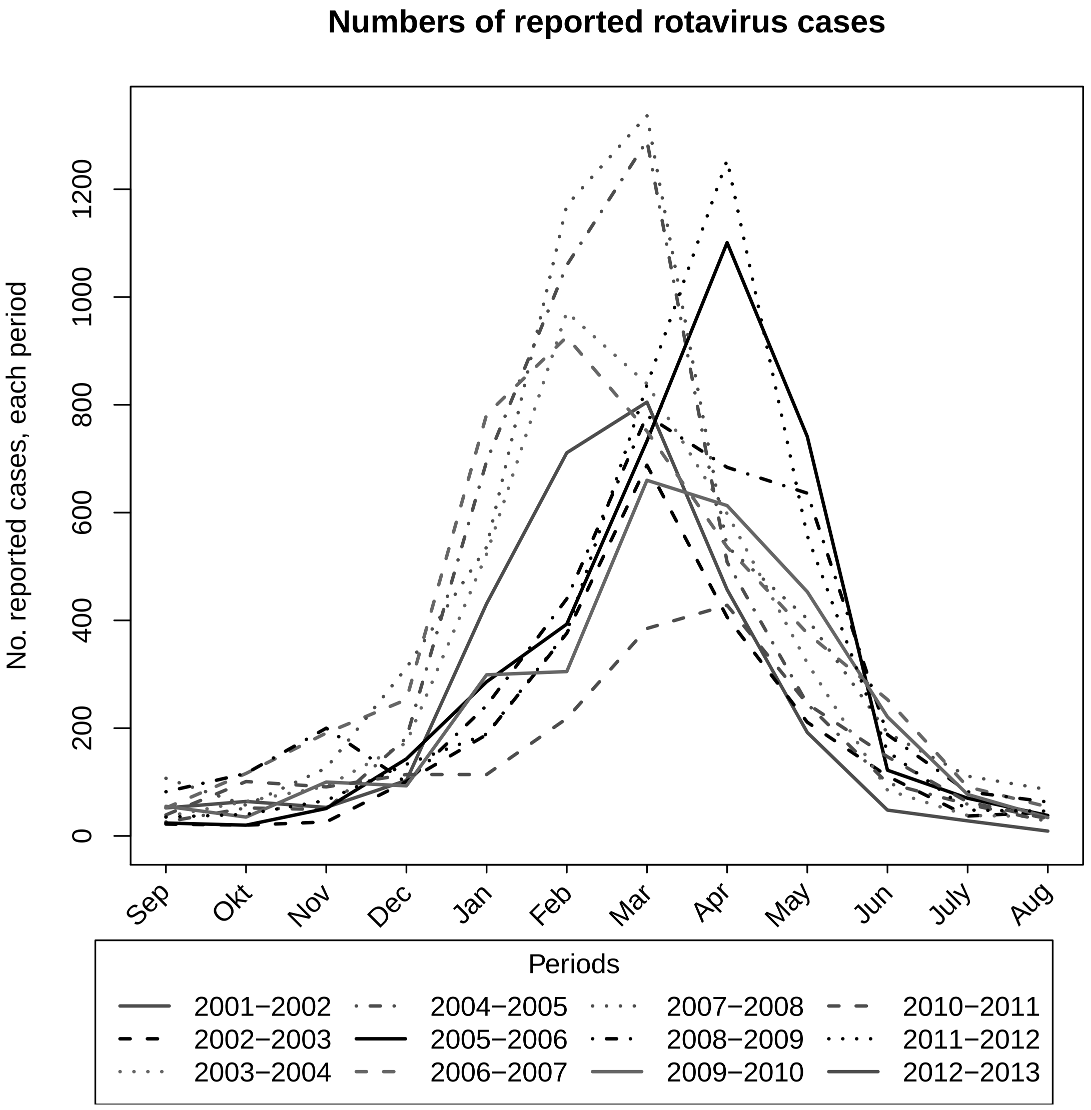

Figure 1.2: Subsequent periods of September to August. In our dataset, there is great seasonal variation.

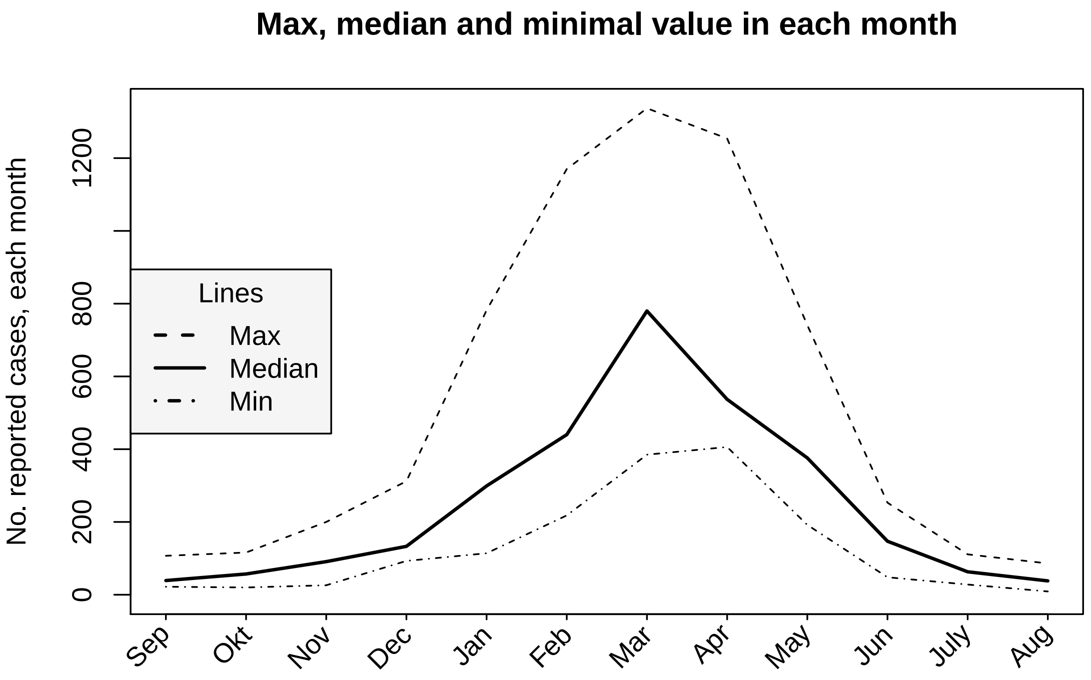

Figure 1.3: Min max and meadian aggregated over months

In Figure 1.2, most cases are located in the period from December to June. Once more, the three last periods from 2009-2013 contain less reported cases with the exception of a peak in April, 2013. The seasonal variation is strong. Comparing Figure 1.2 with Figure 1.3 we can see that periods deviate heavily from the median. In Figure 1.3, the median is generally closer to the smallest value of reported cases and in off-season (July-November) there are less reported cases with a small range.

There are 15 age-categories. In Figure 1.4 the age-categories are shown in a box-plot. The left box-plot shows the inital 15 age-categories while as the right box-plot shows a reduced amount of age-categories. There are 144 observations in each age-category.

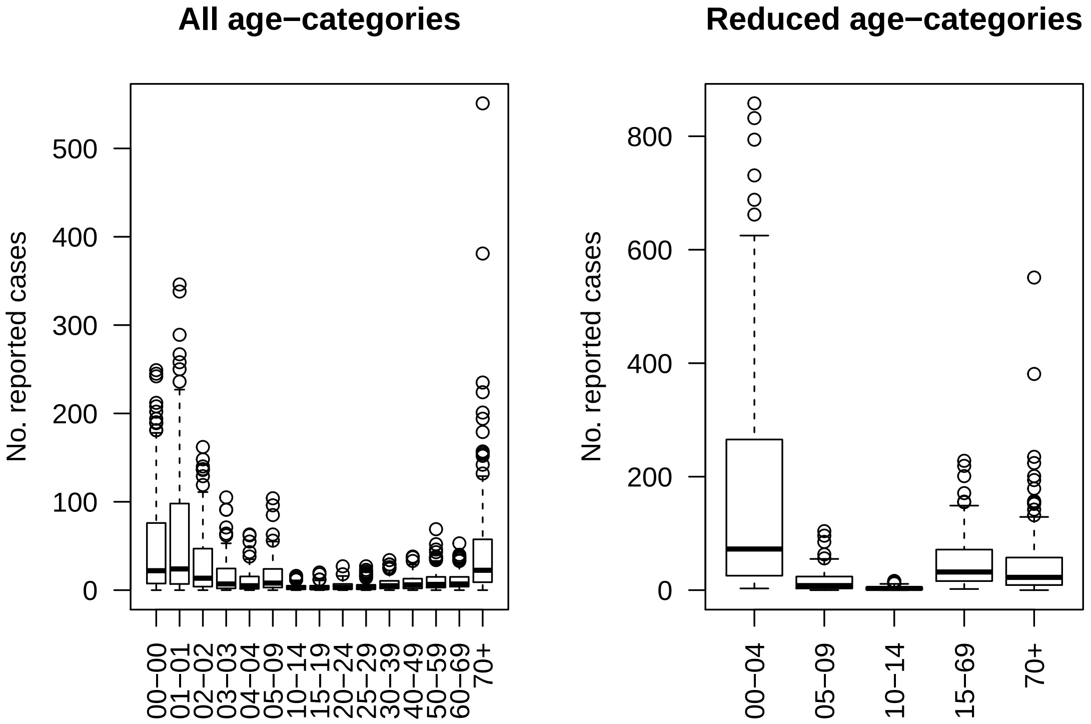

Figure 1.4: Boxplots of age-categories in data. (left) all age-categories (right) reduced amount of age-categories *Note, age-categories do not have the same width.*

Note that in Fig 1.4 the groups do not have the same width. With this in mind, the reported cases in age-categories between 10 to 69 are not many. Most reported cases are in the categories ranging from ”00-04” and age-category ”70+” years. These groups have great spread. In the second (right) box-plot, data is aggregated to five age-categories. These seems to be sufficiently informative for our modeling purpose. The age-groups of most interest was young children, infants and elderly. Thus, the reduced amount of age-categories will not limit the analysis of an age-shift in reported cases. In Fig. 1.1 we saw a decrease in the total reported cases after 2009 followed by a increase in 2013. The recommendation of the vaccine only included children and infants. By looking upon the proportions, we may see if other age-categories may be affected by a decrease in age-category ”00-04”. It follows from the proportions characheristics, a decrease in a age-category implies a increase in another. In Fig 1.5 the proportions of the reduced amount of age-categories are compared.

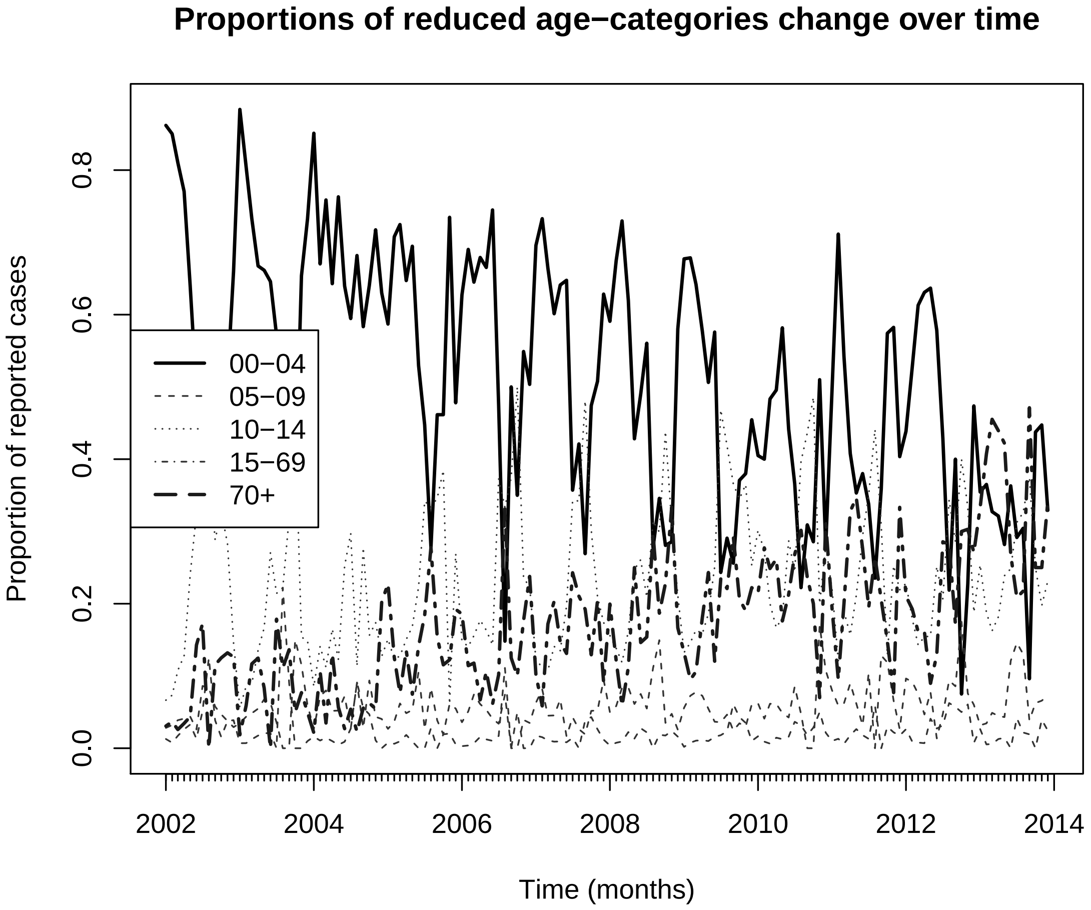

Figure 1.5: Proportions of reduced age-categories. Notice the age-shift betweeen age-category "00-04" and "70+".

In Figure 1.5, proportions of age-categories "05-09" and "10-14" are seen to have no change over time. The increase seen in age-category "15-69" is most likely because of the width of the category. There is a age-shift in proportions after the introduction of the recommendation of the vaccine. The decrease in age-category "00-04" is followed by an increase in the age-category "70+".

We have seen that the recommendation of the vaccine in 2009 caused a decrease in the total reported rotavirus cases, with the exception of a increase in 2013, Fig. 1.1. Most cases where reported in the season December to June. There was strong seasonal variation and some years peaks came later than usual. Using the reduced age-categories, presented in Fig. 1.4, we showed the proportions change over time in Fig. 1.5. The proportions of age-categories showed a age-shift between age-category "00-04" to "70+".

To model the proportions we need to define the multinomial logit- and the Dirichlet-multinomial regression model. The next two chapters are used to derive the two regression models.

# Chapter 2

# Multinomial model

The multinomial distribution is the extension of the binomial distribution. It models nominal data where an outcome has more than two categories. For our application the multinomial distribution will be used for modeling the age-category counts. Each multinomial trial will denote a individual who's been infected by the rotavirus. The parameter $\pi_j$ then represent the probability that a infected individual is of age $j =$(”00-04”,”05-09”,”10-14”,”15-69”,”70+”).

## 2.1 Multinomial distribution

Theory and notation in this section are taken from Agresti (2013, page 6). The vector of cellcounts $\boldsymbol{n}$ contains the counts in each category, i.e. $\boldsymbol{n} = (n_1, n_2, ..., n_J)$ where $n_j$ is the count of trials occured in category $1 \leq j \leq J$. Let $\pi_j = P(Y_{ij} = 1)$ denote the probability of the $i$´th multinomial trial with outcome in category $1 \leq j \leq J$ and let the parameter vector be $\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_J)$. Then

$$\boldsymbol{n}|\boldsymbol{\pi} \sim \text{Multinom}(N; \boldsymbol{\pi})$$

which implies that the probability mass function is

$$f(n_1, n_2, ..., n_J; \boldsymbol{\pi}) = \left(\frac{N!}{\prod_{i=1}^{J} n_i!}\right) \prod_{i=1}^{J} \pi_i^{n_i}, \tag{2.1.1}$$

where $N = \sum_i n_i$, $n_J = N - \sum_i^{J-1} n_i$ and $\pi_J = 1 - \sum_i^{J-1} \pi_i$. For a multinomial distribution and $1 \leq j, k \leq J$ where $j \neq k$

$$\mu_j = \text{E}[n_j] = N\pi_j \tag{2.1.2}$$

$$\text{Var}(n_j) = N\pi_j(1 - \pi_j) = \mu_j \left(1 - \frac{\mu_j}{N}\right) \tag{2.1.3}$$

$$\text{Cov}(n_j, n_k) = -N\pi_j\pi_k = -\frac{\mu_j\mu_k}{N} \tag{2.1.4}$$

Note that both variance (2.1.3) and covariance (2.1.4) have a structure which is given explicitly by the category probabilities, i.e. the expectations. There is no extra parameter which can offset the structure. The parameters $\pi_j, \pi_k$ and the sum of counts $N$ are greater than zero and therefore the covariance is always negative.

The multinomial distribution belongs to a exponential family with natural parameters

$$\boldsymbol{\eta}(\boldsymbol{\pi}) = (\eta_1(\pi_1), ..., \eta_{J-1}(\pi_{J-1})) = (\log \frac{\pi_1}{1 - \sum_{k=1}^{J-1} \pi_k}, ..., \log \frac{\pi_{J-1}}{1 - \sum_{k=1}^{J-1} \pi_k}) \tag{2.1.5}$$

and canonical statistics

$$\mathrm{T}(\boldsymbol{n}) = (\mathrm{T}_1(n_1), ..., \mathrm{T}_{J-1}(n_{J-1})) = (n_1, ..., n_{J-1}).$$

The statistics are sufficient and complete according to Liero and Zwanzig (2012, Corollary 4.1, page 108). The maximum likelihood estimator (MLE) depends on data through the statistics and if there exists a unbiased estimator for the parameters, it is, almost surely, equal to the MLE estimator (Liero and Zwanzig, 2012, Theorem 4.4, page 103).

## 2.2 Multinomial logit regression

The multinomial distribution belongs to a exponential family. This implies that theory for multivariate generalized linear models apply (Fahrmeir and Tutz, 2001). With maximum likelihood estimation (MLE) we fit the multinomial logit regression with the canonical (logit) link. The **MGLM** package (Zhang and Zhou, 2014) for **R**, which will be used for fitting the multinomial logit regression, uses Newton-Raphson iteration to find $\hat{\boldsymbol{\beta}}_{MLE} = \operatorname{argmax}_{\beta} L(\boldsymbol{\beta}|x)$.

The multinomial logit regression model pairs each response category with a baseline category. The baseline category can be chosen arbitrary. In Agresti (2013) the last category or the most common category is suggested to be used as the baseline. For further computations category $J$ will be chosen as the baseline category. Let $\boldsymbol{x}_i = (x_{i0}, ..., x_{ip})^T$ denote the explanatory variables for subject $1 \leq i \leq n$ and $\boldsymbol{\beta}_j = (\beta_{j0}, ..., \beta_{jp})$, $1 \leq j \leq J-1$ , a row vector, as the regression parameters for the j´th baseline-category logit. The first element in the parameter vector $\beta_{j0}$ represents the intercept. Let $\boldsymbol{y}_i = (y_{i1}, ..., y_{iJ})$ represent a multinomial trial for subject $1 \leq i \leq n$. The trial $y_{ij}$ is equal to one whenever a trial occurs in category $j$. The trial may only appear in on category, i.e. $\sum_{j=1}^{J} y_{ij} = 1$. Let $\pi_j(\boldsymbol{x}_i) = \mathrm{P}(y_{ij} = 1|\boldsymbol{x}_i)$ be the probability that the $i$´th trial occurs in category $j$ given a set of covariates $\boldsymbol{x}_i$. The multinomial logit regression model

$$\ln \frac{\pi_j(\boldsymbol{x}_i)}{\pi_J(\boldsymbol{x}_i)} = \boldsymbol{\beta}_j^T \boldsymbol{x}_i \qquad \text{for } j = 1, ..., J-1. \tag{2.2.1}$$

With the logit link we can interpret the coefficients. The exponential of the coefficient, $\exp(\beta_j)$, represents the odds of a trial falling into the category $j$ against category $J$, all other things equal. A odds greater than one represents that a trial is more likely to occur in category $j$ than $J$ and by symmetry if it is less than one it is more likely to occur in category $J$ than $j$. If the odds is equal to one, there is independence between $y$ and covariates.

Using the logit link we have response probabilities

$$\pi_j(\boldsymbol{x}) = \frac{\exp(\boldsymbol{\beta}_j^T \boldsymbol{x})}{1 + \sum_{h=1}^{J-1} \exp(\boldsymbol{\beta}_h^T \boldsymbol{x})}. \tag{2.2.2}$$

In order to fit the multinomial logit regression model by Newton-Raphson iteration we need to derive the log likelihood for regression parameters. For the log likelihood of the regression parameters we use notation from Agresti (2013, section 8.1.4). The likelihood of the regression coefficients is derived from the multinomial likelihood function. For $n$ independent observations

$$L(\boldsymbol{\beta}_i|\boldsymbol{y}_i) = \prod_{i=1}^{n} \prod_{j=1}^{J} \pi_j(\boldsymbol{x}_i)^{y_{ij}}$$

and log likelihood

$$\ell(\boldsymbol{\beta}_i|\boldsymbol{y}_i) = \sum_{i=1}^{n}\sum_{j=1}^{J} y_{ij}\log \pi_k(\boldsymbol{x}_i) = \sum_{i=1}^{n}\sum_{j=1}^{J-1} y_{ij}\log \frac{\pi_j(\boldsymbol{x}_i)}{1-\sum_{k=1}^{J-1}\pi_k(\boldsymbol{x}_i)} + \log\left(1-\sum_{k=1}^{J-1}\pi_k(\boldsymbol{x}_i)\right). \tag{2.2.3}$$

Substituting the response probabilities (2.2.2) into equation (2.2.3) results in the log likelihood for our regression parameters,

$$\sum_{j=1}^{J-1}\left(\beta_{j0}\left(\sum_{i=1}^{n} x_{i0}y_{ij}\right) + \sum_{k=1}^{p}\beta_{jk}\left(\sum_{i=1}^{n} x_{ik}y_{ij}\right)\right) - \sum_{i=1}^{n}\log\left(1+\sum_{j=1}^{J-1}\exp(\boldsymbol{\beta}_j\boldsymbol{x}_i)\right) \tag{2.2.4}$$

Calculations can be found in Agresti (2013, page 298). Sufficient statistics for $\beta_{jk}$ are $\sum_{i=1}^{n} x_{ik}y_{ij}$ for $j = 1,..,J-1$ and $k = 1,...,p$. For $\beta_{j0}$, the intercept, sufficient statistic is equal to $\sum_{i=1}^{n} x_{i0}y_{ij}$ With the canonical link, the fundamental result for regression parameters "For GLM with canonical link, the likelihood equations equate the sufficient statistics for the model paramterers to there expected value". The log likelihood is concave and therefore Newton-Raphson iteration will converge to the MLE.

The Newton-Raphson method is derived in Agresti (2013, page 143) where the following iterative procedure, going from the current value $\boldsymbol{\beta}^t$ to $\boldsymbol{\beta}^{t+1}$, is found

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t - (\boldsymbol{H}^t)^{-1}\boldsymbol{u}^t$$

The superscript $t$ in both $\boldsymbol{H}^t$ and $\boldsymbol{u}^t$ indicates that the matrix and vector is evaluated in $\boldsymbol{\beta}^t$. Iteration starts with a initial guess of regression parameters $\boldsymbol{\beta}^1$. Iteration proceeds until the relative change in the loglikelihoods for two subsequent iterates is less than some constant greater than zero, i.e until the criterion $|\ell(\boldsymbol{\beta}^{t+1}) - \ell(\boldsymbol{\beta}^t)| < \epsilon$ is fulfilled, for some $\epsilon > 0$. The numerical procedure converges fast, relatively few iterations are needed in order to suffice the stopping criterion (Agresti, 2013, page 143-144). The score vector $\boldsymbol{u}$ has elements

$$\boldsymbol{u}^T = \left(\frac{\partial\ell(\boldsymbol{\beta})}{\partial\beta_0}, ..., \frac{\partial\ell(\boldsymbol{\beta})}{\partial\beta_p}\right)$$

and the Hessian matrix $\boldsymbol{H}$ has elements

$$h_{jk} = \frac{\partial^2\ell(\boldsymbol{\beta})}{\partial\beta_j\partial\beta_k}.$$

# Chapter 3

# Dirichlet-multinomial Model

The Dirichlet-multinomial distribution can also be used to model category counts. The most interesting difference to the multinomial distribution is the Dirichlet-multinomial distributions variance-covariance structure. The distributions variance-covariance structure is more complex and is not explicitly defined by the expectations. The structure will have large implications on the fit of the Dirichlet-multinomial regression model and its predictive ability.

In order to derive the Dirichlet-multinomial distribution and regression model, theory regarding the Dirichlet distribution is presented below.

## 3.1 Dirichlet distribution

The Dirichlet distribution is the extension of the beta distribution to more than two categories. In Bayesian statistics, inference for a binomial parameter $\pi$ typically uses the beta distribution as the prior distribution because the beta is the conjugate prior (Agresti, 2013, page 24). The conjugate prior gives us the possibility to express prior knowledge on parameters in terms of mean and variance for the parameter $\pi$. To clarify let

$$\pi \sim \mathrm{Beta}(\alpha_1, \alpha_2)$$

$$y|\pi \sim \mathrm{Bin}(n, \pi).$$

Using Bayes formula (Agresti, 2013, page 23) to update the prior distribution of $\pi$

$$\pi|y \sim \mathrm{Beta}(y + \alpha_1, n - y + \alpha_2).$$

With the same argument as with binomial and beta distribution, one uses the Dirichlet distribution for representing prior knowledge for the multinomial parameter vector $\boldsymbol{\pi}$ (Agresti, 2013, page 25).

Let $\Gamma(x) = (n-1)!$ be the gamma function. With notations taken from Wang Ng. et al. (2011), the Dirichlet distribution with $J$ hyperparameters $\alpha$ has probability density function

$$f(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^{J} \alpha_i)}{\prod_{i=1}^{J} \Gamma(\alpha_i)} \prod_{i=1}^{J} \pi_i^{\alpha_i - 1},$$

where $\boldsymbol{\pi} = (\pi_1, ..., \pi_J)$, $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_J)$, $\alpha_j > 0 \;\; 1 \leq j \leq J$, $\sum_{i=1}^{J} \pi_i = 1$ and the simplex

$$\Omega = \left\{ \boldsymbol{\pi}; \pi_j \in (0,1),\; j = 1, ..., J;\; \sum_{j=1}^{J} \pi_j = 1 \right\} \tag{3.1.1}$$

represents its sample space. As before, the $J\,'$th parameter $\pi_J$ is redundant, $\pi_J = 1 - \sum_{k=1}^{J-1} \pi_k$.

The Dirichlet distribution belongs to the exponential family. The Dirichlet distribution has canonical (or natural) parameters $\boldsymbol{\eta}(\boldsymbol{\alpha}) = (\eta(\alpha_1), ..., \eta(\alpha_J)) = (\alpha_1, ..., \alpha_J)$ and canonical statistics $\boldsymbol{T}(\boldsymbol{\pi}) = (\ln \pi_1, ..., \ln \pi_J)$.

For the Dirichlet distribution on has for $1 \leq i, j \leq J$ with $i \neq j$ (Forbes et al., 2011, page 77)

$$\mathrm{E}[\pi_j] = \frac{\alpha_j}{\sum_{k=1}^{J} \alpha_k} \tag{3.1.2}$$

$$\mathrm{Var}(\pi_j) = \frac{\alpha_j (\sum_{k=1}^{J} \alpha_k - \alpha_j)}{(\sum_{k=1}^{J} \alpha_k)^2 (1 + \sum_{k=1}^{J} \alpha_k)} \tag{3.1.3}$$

$$\mathrm{Cov}(\pi_i, \pi_j) = \frac{-\alpha_i \alpha_j}{(\sum_{k=1}^{J} \alpha_k)^2 (1 + \sum_{k=1}^{J} \alpha_k)}. \tag{3.1.4}$$

The Dirichlet distributions mean, variance and covariance will be used for computations of the Dirichlet-multinomial moments. They are not interesting themselves and therefore no computations will be shown.

## 3.2 Dirichlet-multinomial distribution

When nominal data shows a lot of variance, the multinomial distribution will not be a adequate model category counts because of its limited variance-covariance structure. In search of a more adequate distribution that can model nominal data, one can let $\boldsymbol{\pi}$ be a random variable which follows the Dirichlet distribution. The Dirichlet-multinomial is achieved through integrating the product of the multinomial and dirichlet distribution over the simplex $\Omega$ (3.1.1).

The compound probability mass function for the Dirichlet-multinomial is

$$\begin{aligned} f(\boldsymbol{n}|\boldsymbol{\alpha}) &= \int_{\Omega} f(\boldsymbol{n}|\boldsymbol{\pi}) f(\boldsymbol{\pi}|\boldsymbol{\alpha}) d\boldsymbol{\pi} = \\ &= \frac{N!}{n_1! \cdots n_J!} \frac{\Gamma(\sum_{k=1}^{J} \alpha_k)}{\Gamma(\sum_{k=1}^{J} \alpha_k + n_k)} \prod_{k=1}^{J} \frac{\Gamma(\alpha_k + n_k)}{\Gamma(\alpha_k)} \end{aligned}$$

The Dirichlet-multinomial probability mass function is more thoroughly calculated in appendix section 6.2.1, on page 27. For the Dirichlet-multinomial distribution

$$\mathrm{E}[n_j] = \mu_j = \frac{N \alpha_j}{\sum_{k=1}^{J} \alpha_k} \quad \text{for} \;\; j = 1, ..., J \tag{3.2.1}$$

$$\mathrm{Var}(n_j) = \left( \frac{N + \sum_{k=1}^{J} \alpha_k}{1 + \sum_{k=1}^{J} \alpha_k} \right) \left( \mu_j (1 - \frac{\mu_j}{N}) \right) \quad \text{for} \;\; j = 1, ..., J \tag{3.2.2}$$

$$\mathrm{Cov}(n_j, n_i) = - \left( \frac{N + \sum_{k=1}^{J} \alpha_k}{1 + \sum_{k=1}^{J} \alpha_k} \right) \frac{\mu_j \mu_i}{N} \quad ,\; j \neq i,\; j, i = 1, .., J. \tag{3.2.3}$$

Computations are shown in appendix, page 27. As with the multinomial distribution, covariance is always negative as $\alpha_j > 0$ for $j = 1, ..., J$.

The Dirichlet-multinomial variance-covariance structure is determined by the first moment and the fraction

$$\left(\frac{N+\sum_{k=1}^{J}\alpha_k}{1+\sum_{k=1}^{J}\alpha_k}\right)\geq 1. \tag{3.2.4}$$

In (3.2.4) equality occurs when $N=1$. For a given N greater than one, covariance and variance diminishes when the $\sum_{k=1}^{J}\alpha_k$ grows, which is when the amount of categories grow. Now, compare (3.2.2), (3.1.4) with (2.1.3) and (2.1.4) on page 7. Fraction (3.2.4) can offset the variance-covariance structure where as the multinomial distributions was explicitly formulated by the mean. This flexible structure will have large implications on the Dirichlet-multinomial regression fit.

## 3.3 Dirichlet-Multinomial regression

The theory for multivariate generalized linear models base upon two assumptions. The first assumption is the *distribution assumption*, which states that the distribution must belong to the exponential family (Fahrmeir and Tutz, 2001, page 76). In comparison to the mutlinomial distribution, the Dirichlet-multinomial distribution does belong to the exponential family. The generalized linear model framework will therefore not apply. The regression is still possible through choosing a link function between covariates and distribution parameters, $\boldsymbol{\alpha}$. The choice of different link functions has been left out as it is specified by the **MGLM** package (Zhang and Zhou, 2014). In this package the link function

$$\alpha_j(\boldsymbol{x}_i)=\exp(\boldsymbol{\beta}_j^T\boldsymbol{x}_i), \tag{3.3.1}$$

for $j=1,...,J$, is used, where $\boldsymbol{\beta}_j=(\beta_{j0},...,\beta_{jp})$, $1\leq j\leq J$ the regression parameters and $\boldsymbol{x}_i$ the subset of the design matrix which are defined on page 8. The same suggestion for link function was made by Chen and Li (2013). Using the log link we consider the following model

$$\boldsymbol{\pi}_j(\boldsymbol{x}_i)=\frac{\alpha_j(\boldsymbol{x}_i)}{\sum_{k=1}^{J}\alpha_k(\boldsymbol{x}_i)}=\frac{\exp(\boldsymbol{\beta}_j^T\boldsymbol{x}_i)}{\sum_{k=1}^{J}\exp(\boldsymbol{\beta}_k^T\boldsymbol{x}_i)},\quad j=1,...,J. \tag{3.3.2}$$

In comparison to the multinomial logit regression there is no need to set a baseline category. This implies that we have $p$ more parameters in the Dirichlet-multinomial regression model. The sign of $\beta_{jl}$ determines whether a Dirichlet-multinomial trial is more or less likely to occur in category $j$, given a increase in $x_{il}$, $1\leq l\leq p$. Within the model, the increase in covariate $x_{il}$ has complex implications on the fitted proportions. To show this let $\boldsymbol{x}_i,\tilde{\boldsymbol{x}}_i$ s.t that they only differ as $\tilde{x}_{il}=x_{il}+1$ and consider the following fraction

$$\begin{aligned}\frac{\boldsymbol{\pi}_j(\tilde{\boldsymbol{x}}_i)}{\boldsymbol{\pi}_j(\boldsymbol{x}_i)}&=\frac{\exp(\boldsymbol{\beta}_j^T\tilde{\boldsymbol{x}}_i)}{\sum_{k=1}^{J}\exp(\boldsymbol{\beta}_k^T\tilde{\boldsymbol{x}}_i)}\cdot\frac{\sum_{k=1}^{J}\exp(\boldsymbol{\beta}_k^T\boldsymbol{x}_i)}{\exp(\boldsymbol{\beta}_j^T\boldsymbol{x}_i)}=\\&=\frac{\exp(\beta_{j0}+...+\beta_{jl}(x_{il}+1)+...+\beta_{jp}x_{ip})}{\exp(\beta_{j0}+...+\beta_{jl}x_{il}+...+\beta_{jp}x_{ip})}\cdot\frac{\sum_{k=1}^{J}\exp(\boldsymbol{\beta}_k^T\boldsymbol{x}_i)}{\sum_{k=1}^{J}\exp(\boldsymbol{\beta}_k^T\tilde{\boldsymbol{x}}_i)}=\\&=\exp(\beta_{jl})\cdot\frac{\sum_{k=1}^{J}\exp(\boldsymbol{\beta}_k^T\boldsymbol{x}_i)}{\sum_{k=1}^{J}\exp(\boldsymbol{\beta}_k^T\tilde{\boldsymbol{x}}_i)}.\end{aligned} \tag{3.3.3}$$

The sign of $\beta_{jl}$ determines whether the exponential of the regression parameter is greater or less than one. Also, the fraction in (3.3.3) will limit the contribution of $\exp(\beta_{jl})$ on fitted mean values. We see this by considering the case when $\beta_{jl} > 0$. Then $\exp(\beta_{jl})$ is greater than one and the fraction in (3.3.3) is less than one since

$$\frac{\sum_{k=1}^{J} \exp(\boldsymbol{\beta}_k^T \boldsymbol{x}_i)}{\sum_{k=1}^{J} \exp(\boldsymbol{\beta}_k^T \tilde{\boldsymbol{x}}_i)} < 1 \Leftrightarrow \sum_{k=1}^{J} \exp(\boldsymbol{\beta}_k^T \boldsymbol{x}_i) - \exp(\boldsymbol{\beta}_k^T \tilde{\boldsymbol{x}}_i) < 0 \Leftrightarrow \exp(\boldsymbol{\beta}_j^T \boldsymbol{x}_i)(1 - \exp(\beta_{jl})) < 0 \quad (3.3.4)$$

The results implies that, since $\beta_{jl}$ is positive, fraction (3.3.3) will diminish $\exp(\beta_{jl})$ contribution to fitted mean values. Of course, the converse is also true. Whenever $\beta_{jl}$ is negative, fraction (3.3.3) will amplify $\exp(\beta_{jl})$ contribution.

Now, with $n$ independent observations, kernel of the Dricihlet-multinomial and the suggested link function we have the log likelihood for regression parameters (Chen and Li, 2013)

$$\begin{aligned} \ell(\boldsymbol{\beta}|\boldsymbol{x},\ \boldsymbol{n}) = & \sum_{i=1}^{n} \left[ \log\Gamma\left(\sum_{j=1}^{J} \alpha_j(\boldsymbol{x}_i)\right) - \log\Gamma\left(\sum_{j=1}^{J} n_{ij} + \sum_{j=1}^{J} \alpha_j(\boldsymbol{x}_i)\right)\right. \\ & \left. + \sum_{j=1}^{J} \log\Gamma(n_{ij} + \alpha_j(\boldsymbol{x}_i)) - \log\Gamma(\alpha_j(\boldsymbol{x}_i))\right] \end{aligned} \quad (3.3.5)$$

The gamma-function $\Gamma(\cdot)$ is differentiable and therefore, it should pose no problem in finding the components for Newton-Raphson method since the score vector and the Hessian matrix can be calculated analytically.

# Chapter 4

# Data analysis of Rotavirus data

The recommendation of the vaccine in 2009 seemed to have caused a decrease in the reported cases among young children and infants. With the reduced amount of age-groups, presented in Figure 1.4 on page 5, we have five different age-groups which will be used in the fitting procedure.

When fitting, we use the covariates time (t) and two harmonic components $\sin(\frac{2\pi}{12} \cdot t)$ and $\cos(\frac{2\pi}{12} \cdot t)$, in order to capture periodicity. Figure 1.2 on page 3, showed strong seasonal variation. The two harmonic components will not be able to capture this seasonal variation but no further attention will be put onto variable selection as it is not included in the aim of this thesis.

The design matrix has the following appearance , $\boldsymbol{x}_t = (1, t, \sin(\frac{2\pi}{12} \cdot t), \cos(\frac{2\pi}{12} \cdot t))$, $1 \leq t \leq 144$, and parameter vector $\boldsymbol{\beta}_j = (\beta_{j0}, \beta_{j1}, \beta_{j2}, \beta_{j3})$ where $\beta_{j0}$ represents the intercept. The linear covariate will set the primary development of the fitted mean values as the harmonic components are periodic, taking values in the interval $[-1, 1]$.

## 4.1 Data analysis using the multinomial logit regression

The multinomial logit regression model sets one category to the baseline category. With the last category as the baseline, i.e. category ”70+”, we have the multinomial logit regression model

$$\log\left(\frac{\pi_j(\boldsymbol{x}_t)}{\pi_{70+}(\boldsymbol{x}_t)}\right) = \beta_{j0} + \beta_{j1} t + \beta_{j2} \sin\left(\frac{2\pi}{12} \cdot t\right) + \beta_{j3} \cos\left(\frac{2\pi}{12} \cdot t\right), \tag{4.1.1}$$

where j=(”00-04”,”05-09”,”10-14”,”15-69”). All interpretation of regression coefficient and models are done in comparison to the ”70+” category.

Using the **MGLM** package (Zhang and Zhou, 2014) we fit the multinomial logit regression, as follows

```r
Multinom_mod <- MGLMreg(as.matrix(RotaVirusBB[, agNames]) ~
    t + sin + cos, data = RotaVirusBB, dist = "MN")
```

In Table 4.1, estimated regression parameters for the linear covariate, $t$, are all negative. As time grows there is a multiplicative reduction in the odds by

$$((e^{\hat{\beta}_{1,1} \cdot t}, e^{\hat{\beta}_{1,2} \cdot t}, e^{\hat{\beta}_{1,3} \cdot t}, e^{\hat{\beta}_{1,4} \cdot t}) = (0.9823^t, 0.9892^t, 0.9861^t, 0.9907^t)$$

| | 00-04 | 05-09 | 10-14 | 15-69 | wald value | Pr(>wald) |
|---|---|---|---|---|---|---|
| (Intercept) | 2.463 | -0.239 | -1.140 | 0.998 | 6921.434 | 0.000 |
| t | -0.018 | -0.011 | -0.014 | -0.009 | 2296.300 | 0.000 |
| sin | 0.269 | 0.081 | -0.607 | -0.390 | 786.837 | 0.000 |
| cos | 0.115 | 0.011 | -0.209 | -0.046 | 83.688 | 0.000 |

Table 4.1: Multinomial regression parameter estimate and Wald tests for each parameter.

for the four categories. The trend, among all categories, is that a rotavirus case becomes less likely to occur in the age-categories ”00-04”, ”05-09”, ”10-14” and ”15-69” as time progress. Thus it becomes more likely to occur in the age-category ”70+”. In Table 4.1, intercept of age-categories ”05-09” and ”10-14” is negative. All other things equal, the odds development over time shows that it is more likely that a rotavirus case is to occur in age-category ”70+” than in ”05-09” or ”10-14”. The effects of the linear covariate are small compared to the harmonic components. As the rotavirus was most common among infants and young children (Koch and Wiese-Posselt, 2011), the age-category of most interest is ”00-04”. Consider the following model

$$\log\left(\frac{\boldsymbol{\pi}_{00-04}(\boldsymbol{x}_t)}{\boldsymbol{\pi}_{70+}(\boldsymbol{x}_t)}\right) = 2.4627 - 0.0179t + 0.2692\sin\left(\frac{2\pi}{12}\cdot t\right) + 0.1148\cos\left(\frac{2\pi}{12}\cdot t\right). \quad (4.1.2)$$

The odds of a reported case occuring in category ”00-04” instead of ”70+” is influenced by all covariates. As previously described, the linear covariate will determine the trend and the two harmonic covariates will show seasonality. In Figure 4.1 the odds development over time are shown, with and without harmonic covariates.

```r
beta_hat <- Multinom_mod$coeff
Design_mat <- t(Multinom_mod$data$X)
Odds <- exp(beta_hat[, 1] %*% Design_mat)  # choose category '00-04'
```

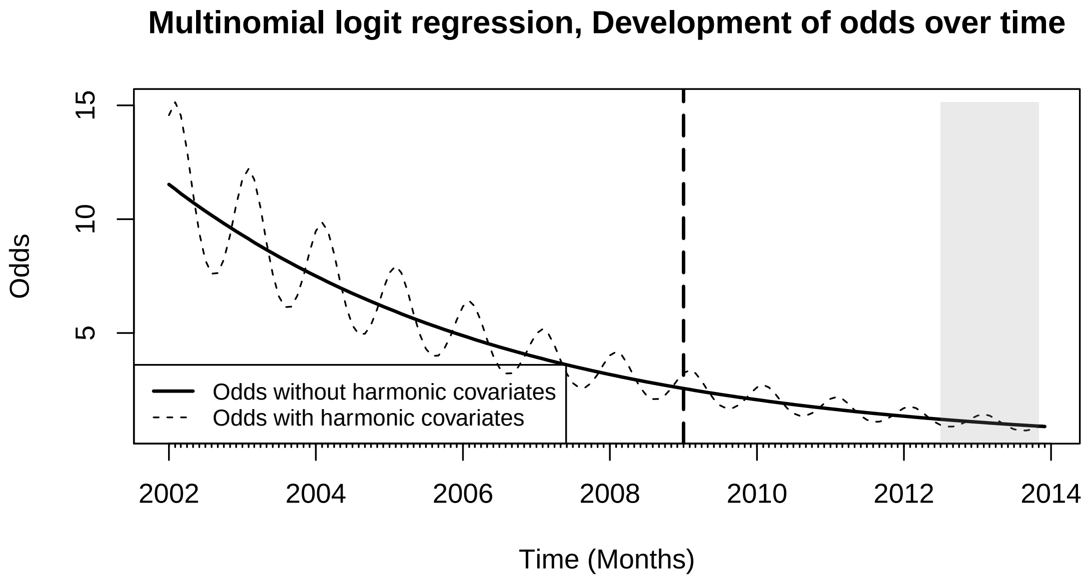

Figure 4.1: Odds for age category ”00-04” plotted over time, grey area indicates values of odds less than one

Once again, the vertical line in 4.1 indicates the recommendation of the vaccine. In 2012-07-01 the odds attain values below one. At this moment in time a reported virus case is more likely to appear in age-category ”70+” than in category ”00-04”.

## 4.2 Data analysis using the Dirichlet-multinomial regression

In section 3.3, we saw that the Dirichlet-multinomial regression model did not need to pursue a baseline category. Therefore, we have four more regression parameters in the following model

$$\boldsymbol{\pi}_j(\boldsymbol{x}_t) = \frac{\exp(\boldsymbol{\beta}_j^T \boldsymbol{x}_t)}{\sum_{k=1}^{5} \exp(\boldsymbol{\beta}_k^T \boldsymbol{x}_t)}, \tag{4.2.1}$$

where j=(”00-04”, ”05-09”, ”10-14”, ”15-69”,”70+”). We fit the Dirichlet-multinomial regression using the **MGLM** package

```
Dirichlet_mod <- MGLMreg(as.matrix(RotaVirusBB[, agNames]) ~
    t + sin + cos, data = RotaVirusBB, dist = "DM")
```

| | 00-04 | 05-09 | 10-14 | 15-69 | 70+ | wald value | Pr(>wald) |
|---|---|---|---|---|---|---|---|
| (Intercept) | 4.380 | 1.731 | 0.910 | 3.010 | 2.038 | 1936.640 | 0.000 |
| t | -0.012 | -0.004 | -0.006 | -0.004 | 0.005 | 363.720 | 0.000 |
| sin | 0.542 | 0.304 | -0.243 | -0.169 | 0.172 | 261.025 | 0.000 |
| cos | -0.072 | -0.156 | -0.236 | -0.257 | -0.150 | 20.932 | 0.001 |

Table 4.2: Estimated coefficients of Dirichlet-multinomial model and Wald tests for regression parameters

In Table 4.2, for age-category ”70+”, the sign of the coefficient representing the linear covariate, is positive. As time $t$ progress, it is more likely that a reported case occurs in age-category ”70+”. The two regression models show the same trend over time. As time progress it becomes more likely that rotavirus cases appear in category ”70+” and less likely that it will appear in any other age-category.

The two regression models do not use the same link function so the estimates can not be compared directly. Using **R**’s *predict* function, one can calculate the fitted mean values for the regression models.

In Figures 4.2-4.4 the fitted mean values are dispalyed, i.e. $\hat{\pi}_{jt}$ for $1 \leq t \leq 144$ where $j =$(”00-04”,”05-09”,”10-14”,”15-69”,”70+”).

```
pred_Multinom <- predict(Multinom_mod, newdata = t(Design_mat))
pred_dirich <- predict(Dirichlet_mod, newdata = t(Design_mat))
```

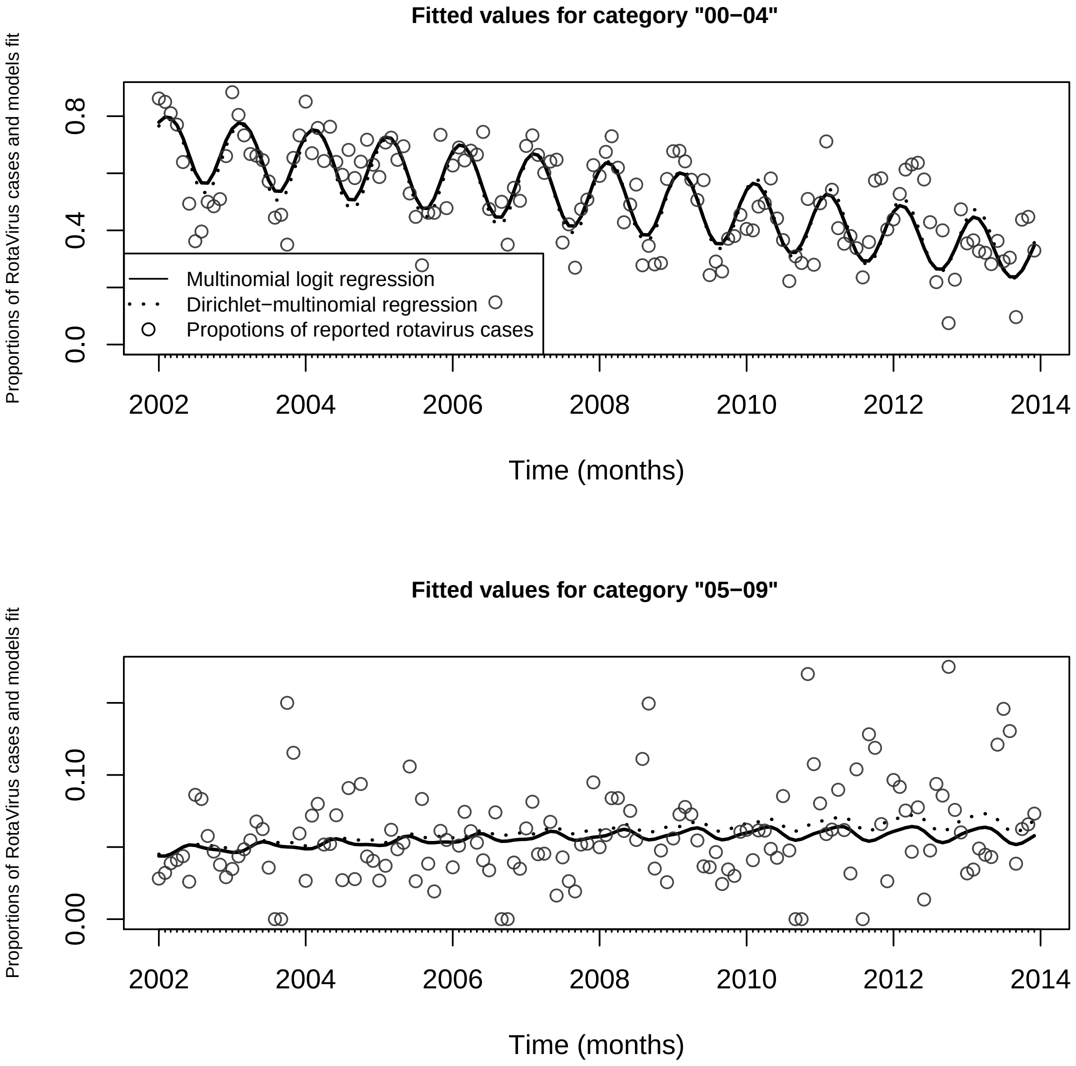

Figure 4.2: Fitted mean values age-categories ”00-04”, ”05-09” ”10-14”, ”15-69” and ”70+”

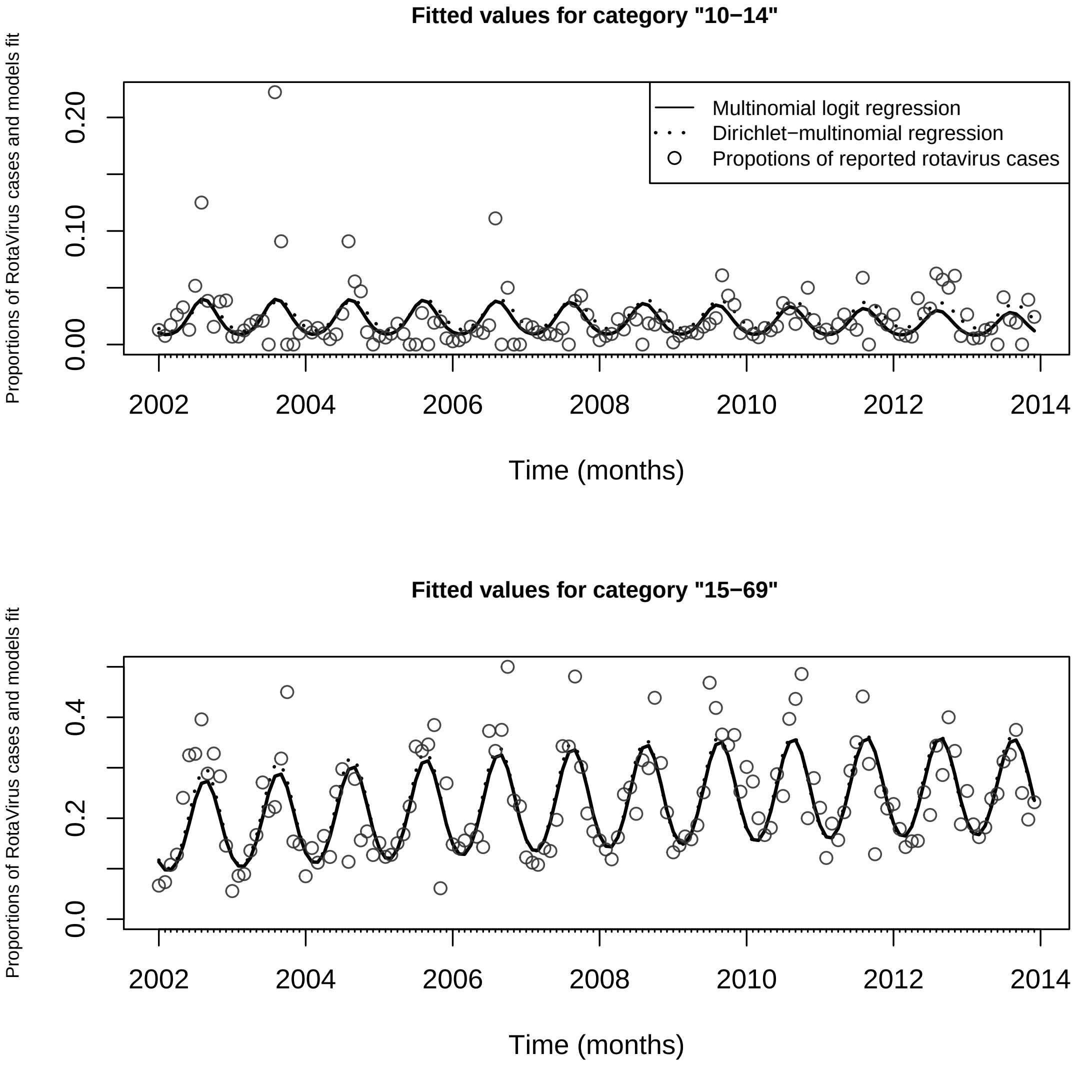

Figure 4.3: Fitted mean values age-categories ”00-04”, ”05-09” ”10-14”, ”15-69” and ”70+”

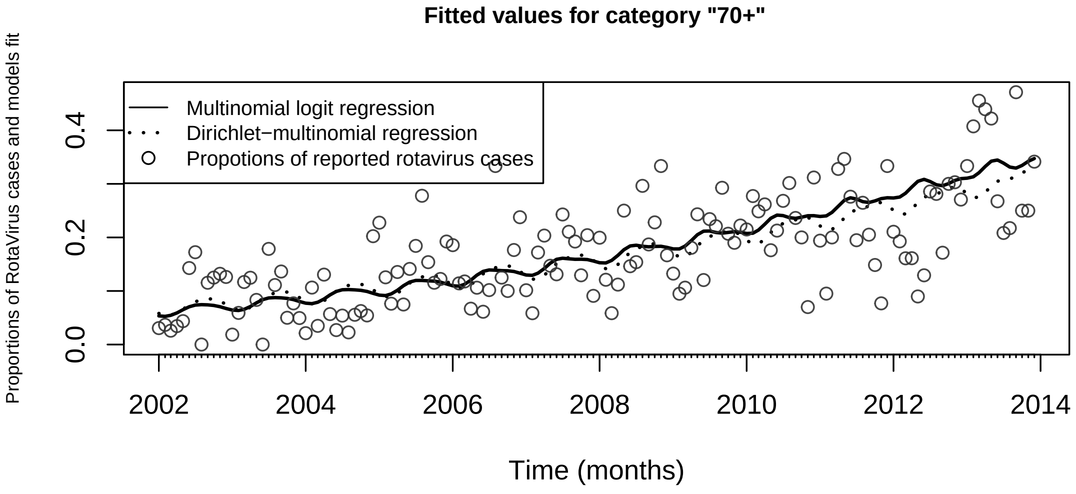

Figure 4.4: Fitted mean values age-categories ”00-04”, ”05-09” ”10-14”, ”15-69” and ”70+”

There are minor differences in fitted values in all age-categories. Both models suggests a increase in the probability that a virus case will appear in category ”70+”. For this age-category the Dirichlet-multinomial suggest a smaller increase in the fitted mean values over time. The fitted values of age-categories ”05-09” and ”10-14” are small. The probability of a rotavirus case occuring in these age-categories are small throughout the whole period. The fitted values of the wide age-category ”15-69” takes on values of almost equal size to the fitted values of age-category ”70+”. One could suggest that it is the width of the age-category ”15-69” which is causing the large fitted values.

From this graphical presentation there little to say of which model fits data best. The goodness-of-fit measures AIC (*Aikaikes information critera*) and BIC (*Bayesian information criterion*) should give a indication on which model fits mean values more adequate. The two measures are defined as

$$AIC = G^2 - 2 \cdot \mathrm{df}$$
$$BIC = G^2 - \log(n) \cdot (\mathrm{df})$$

where $G^2 = -2 \log \Lambda$ is the likelihood-ratio statistic (Agresti, 2013, page 212-213).

| | Multinomial regression | Dirichlet-multinomial regression |
|---|---|---|
| AIC | 4836.43 | 3691.10 |
| BIC | 4883.95 | 3750.49 |

Table 4.3: Goodness-of-fit measures for the two regression models representing a tradeoff between fit and model complexity

Both AIC and BIC represents a tradeoff between obtaining a good model while penalizing for the complexity of the model. AIC penalizes with two times the number of degrees of freedom (residual df or number of parameters) while BIC penalizes by $\log(n)$ times the residual df. As an example the Dirichlet-multinomial will be penalized by 8 more units (AIC). In Table 4.3 both AIC and BIC are approximately a thousand units lower for the Dirichlet-multinomial regression model. This implies that the Dirichlet-multinomial fitted mean values are closer to the mean values given data (Agresti, 2013, page 212).

The models do achieve large differences in modeling the mean, confirmed in table 4.3. As a result of the different variance-covariance structure there should also be large differences in predicitve intervals (PI). Since it is not possible to find analytic closed form expressions for the pred. distributions, PI's are constructed by sampling. Using the asymptotic multivariate distribution for the maximum likelihood estimator (Hoadley, 1971) we can create a sampling algorithm. Once again, let $j$ representing the age-category. The function *MGLMpredInt* is constructed so that it will implement the algorithm once. Using the **R** function *replicate* we can sample repeated times. The PI quantiles are found using the **R** function *quantile*.

| | Sampling algorithm to obtain predictive distribution |
|---|---|
| 1. | Sample $\boldsymbol{\beta} \sim \text{MVN}(\hat{\boldsymbol{\beta}}, (-\boldsymbol{H})^{-1})$ |
| 2. | Calculate distribution parameters, i.e. $\boldsymbol{\pi}(\boldsymbol{\beta})$ or $\boldsymbol{\alpha}(\boldsymbol{\beta})$,<br>using their respective regression link ((2.1) or (3.3)) |
| 3. | Using the estimated parameters calculated in step 2 to sample $n_{tj}$, $t = 1, ..., 144$<br>and $j = 1, ..., 5$ from $\text{Multi}(N; \boldsymbol{\pi}(\boldsymbol{\beta}))$ or $\text{Dirichlet-Multi}(N; \boldsymbol{\alpha}(\boldsymbol{\beta}))$ |
| 4. | Calculate the proportions at time $t$, $P = n_{tj} / \sum_{j=1}^{J} n_{tj}$. |

Table 4.4: Sampling algoritm which is implemented by the function *MGLMpredInt*

```
N <- 1000  # Number of replicates.

New_propMultinom <- replicate(N, MGLMpredInt(Multinom_mod))
New_propDirichlet <- replicate(N, MGLMpredInt(Dirichlet_mod))


oneAGPI <- function(ageIndex, predictive) {
    res <- matrix(apply(X = predictive[, ageIndex, ], 1,
        FUN = quantile, probs = c(0.025, 0.975)), ncol = 2,
        byrow = TRUE)
    return(res)
}

ageInd <- c(`00-04` = 1, `05-09` = 2, `10-14` = 3, `15-69` = 4,
    `70+` = 5)
piMulti <- lapply(ageInd, oneAGPI, predictive = New_propMultinom)
piDirMulti <- lapply(ageInd, FUN = oneAGPI, predictive = New_propDirichlet)
```

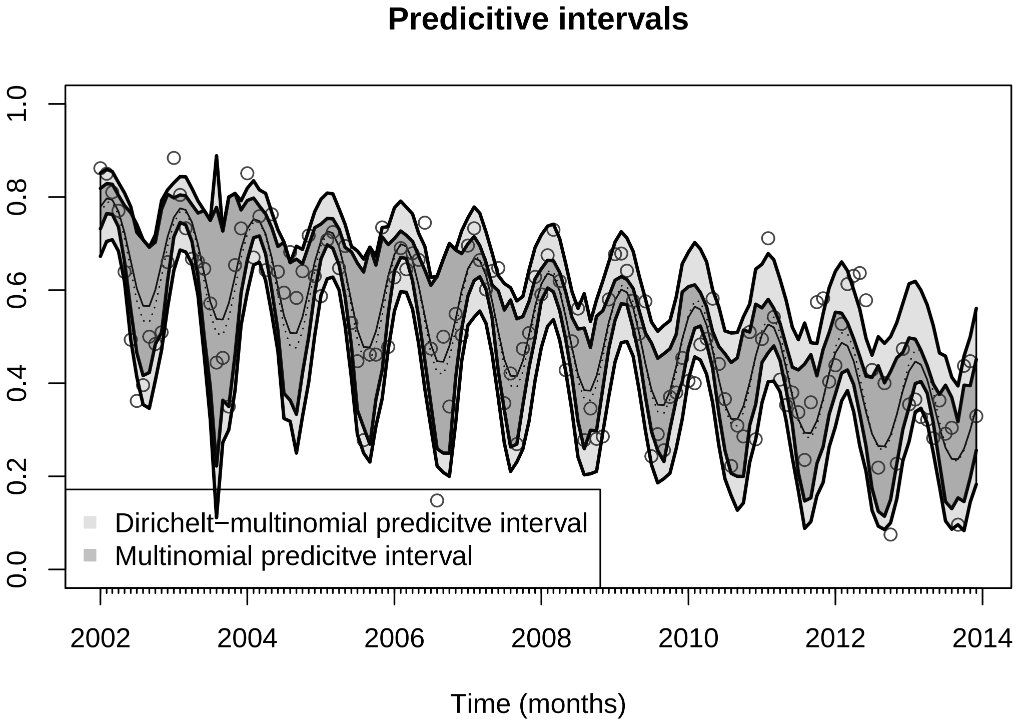

Figure 4.5: 95 % Predicitive intervals for multinomial and Dirichlet-multinomial, age category "00-04"

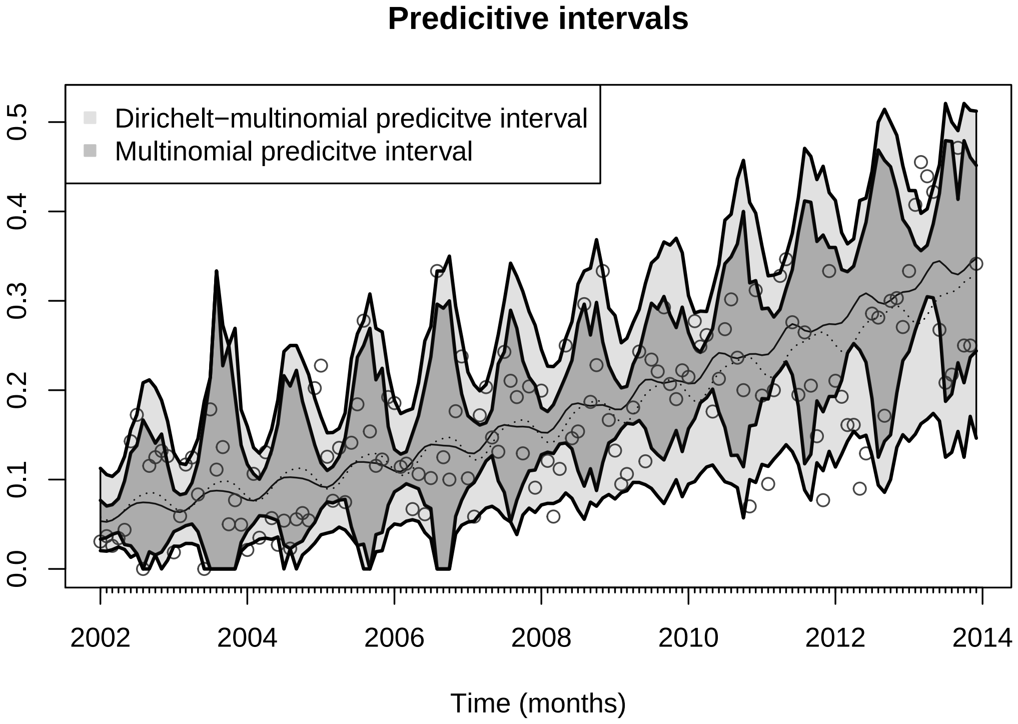

Figure 4.6: 95 % Predictive intervals for multinomial and Dirichlet-multinomial, age category "70+"

Figure 4.5 and 4.6 show two out of five age-categories. In the third step of the sampling algorithm, described in Table 4.4, the sampled counts will inherit the distributions variance-covariance structure. The predictive intervals for the Dirichlet-multinomial will inherit the more flexible variance-covariance structure, described in chapter 3. As seen in Figures 4.5 and 4.6 the predictive interval for the Dirichlet-multinomial is wider and contains more observations. For category ”70+” the Diriclet-multinomial predictive interval includes 0.91% where as the multinomial includes 0.57 % of the observations. For the younger age-category the predicitive interval for the Dirichlet-multinomial contain 0.86 % of the observations and the predictive interval for the multinomial contain 0.54%.

# Chapter 5

# Discussion

The multinomial distribution and the Dirichlet-multinomial distribution (chapter 2 through 3) models nominal data. They presented very different variance-covariance structures. The multinomial distribution had a more limited structure, explicitly defined by the parameters ((2.1.3) and (2.1.4)) while the Dirichlet-multinomial distribution showed a more flexible variance-covariance structure ((3.1.3) and (3.1.4)). In the descriptive data analysis (section 1.1) we concluded that data showed a lot of season variation (fig.1.2, page 3) as well as a age-shift in proportions (fig.1.5, page 6) between age-categories ”00-04” to ”70+”.

With the **MGLM** package (Zhang and Zhou, 2014) for **R** (R Core Team, 2014), we modeled the proportions using the multinomial logit and Dirichlet-multinomial regression in chapter 4. The two categorical regression models used different link functions and therefore could not be compared directly. A graphical presentation of fitted mean values (Fig. 4.2, 4.3 and 4.4) showed minor differences between models. The two goodness-of-fit measures AIC and BIC (table 4.3 page 19) showed results highly in Dirichlet-multinomial regression models favor. AIC and BIC measures how close fitted mean values are to the true mean values, given data, and penalize models with many parameters (Agresti, 2013, page 212). The Dirichlet-multinomial regression did not need to pursue a baseline-category and therefore had 4 more parameters compared to the multinomial logit regression. Thus, the Dirichlet-multinomial will be penalized harder by AIC and BIC. The difference between models AIC and BIC showed that the extra parameters made no difference. AIC and BIC showed a difference around a thousand units less for the Dirichlet-multinomial regression. It is much larger than $2 \cdot 4$ (AIC) which should be the difference if the two regression models fitted mean values alike. The predictive intervals (Fig. 4.5 and 4.6), constructed through sampling, also showed large differences in the two regression models. The predictive interval for the Dirichlet-multinomial regression contained on avarage 90% of the observations. The multinomial predictive interval contained on average 50-60% of the observation. The multinomial logit regression model is not adequate to model proportions of age-categories of reported rotavirus cases of Brandenburg, Germany. The Dirichlet-multinomial, on the other hand, showed much better fit (AIC and BIC) and sampled predictive intervals where acceptable when overdispersion was present in data. The aim was to show the practical use of the Dirichlet-multinomial regression model and that it was more flexible than the multinomial logit regression model, which should be regarded as accomplished with this analysis.

In the data analysis (chapter 4 on page 14) we saw that both regression models described a increase in the fitted mean values for age-category ”70+” as time progress (Fig. 4.4). Fitted mean values for age-category ”00-04” showed a decrease in reported rotavirus cases(Fig. 4.2). The age-shift we saw in Figure 1.5 on page 6, is confirmed by both models. Elderly are affected by the vaccine programme.

The assumption of age-categories being a nominal scale may not affect the model but the interpretation is not correct. A nominal scale has categories which can not be determined as better or worse, where as age-categories of reported rotavirus cases are certainly ordinal. There exist alternative models based on the multinomial logit regression which handle ordinal data. In Agresti (2013, section 8.2) suggest a model which utilize the category ordering by forming logits of cumulative probabilities. These where not of interest because it would limit (if not exclude) the possibility to compare the multinomial logit regression model to the Dirichlet-multinomial regression model as there does not seem to exist any extensions of the Dirichlet-multinomial regression model which handles ordinal data.

## 5.1 Future work

The results may be somewhat incomplete as there was no attention put to variable selection. On the other hand, with this type of analysis, the two regression models where given the same prerequisites and focus was on presenting the mathematics. With some attention put on variable selection we may have improved fitted mean values. It would be of great interest to see how much better fit one would receive with some time spent on variable selection. Also, if relieving the assumption of no auto-correlation, the addition of auto-regressive and moving-average terms may decrease the distance between the two categorical regression models AIC and BIC and predictive intervals. Theory for categorical time series regression models is presented in Fokianos and Kedem (2003), though only for the multinomial setting.

It would also be interesting to model the absolute numbers of rotavirus cases per age-category in order to answer whether the vaccine was effective or not. The nature of proportions implies that if reported cases for one age-category reduces another one has to go up. A interesting modification made in Höhle et al. (2011) where the authors fitted independently a ARMA model for the mean number of varicella cases per sentinel reporting unit (level 1 model) and a multinomial logit regression model proportions of age-categories (level 2 model). They predicted the mean at time $t$ using the level 1 model and then multiplied by the age-category probabilities (level two model) to receive a time-series model for the mean number of cases per sentinel unit in each group. The extension on our application would imply mean number reported rotavirus cases in each age-category. It would also be interesting to see how the Dirichlet-multinomial model could improve the described model.

## Acknowledgements

I would like to thank Michael Höhle for being a great supervisor and educationist in the making of this thesis!

### R-code and functions

Functions and R-code is included in a separate ZIP-file or can otherwise be obtained by contacting me at Ethorsn@gmail.com.

# Chapter 6

# Appendix

The following sections are present in order to clarify and show computations of distributions, mean, variance and covariance structure.

## 6.1 The Multinomial distribution

To calculate expected value, variance and covariance for the multinomial distribution we use the moment generating function defined in Gut (2009, page 63) with theorem 3.3 Gut (2009, page 64).

### 6.1.1 Expected value, variance and covariance

Let $\mathbf{t} = (t_1, t_2, ..., t_{J-1})$ where $J$ is the number of categories, let the countvector $\boldsymbol{n} = (n_1, ..., n_{J-1})$, $n_J = N - \sum_{k=1}^{J-1} n_k$, the parametervector $\boldsymbol{\pi} = (\pi_1, ..., \pi_{J-1})$ and $\pi_J = 1 - \sum_{k=1}^{J-1} \pi_k$. The moment generating function for the multinomial distribution is equal to

$$\psi(\boldsymbol{t}) = E[\exp\{\boldsymbol{t}\boldsymbol{n}^T\}] = \sum_{\boldsymbol{n}} \exp\{\boldsymbol{t}\boldsymbol{n}^T\} f(N; \boldsymbol{n}, \boldsymbol{\pi}) =$$

$$= \sum_{\boldsymbol{n}} exp\{\boldsymbol{t}\boldsymbol{n}^T\} \left( \frac{N!}{\prod_{k=1}^{J} n_k!} \right) \prod_{k=1}^{J} \pi_i^{n_i} =$$

$$= \sum_{\boldsymbol{n}} \left( \frac{N!}{\prod_{i=1}^{J-1} n_i!(N - \sum_{i=1}^{J-1} n_i)!} \right) \prod_{i=1}^{J-1} (e^{t_i}\pi_i)^{n_i} (1 - \sum_{i=1}^{J-1} \pi_i)^{N - \sum_{i=1}^{J-1} n_i} =$$

$$= \left( \pi_1 e^{t_1} + \pi_2 e^{t_2} + ... + \pi_{J-1} e^{t_{J-1}} + 1 - \sum_{k=1}^{J-1} \pi_k \right)^N \tag{6.1.1}$$

In the last step we used the fact that it is a multinomial serie and using the extension of the binomial theorem (*Multinomial Series, From MathWorld–A Wolfram Web Resource*). Using equation (6.1.1) and $\psi^{(n)}_{t_j}(0) = \mathrm{E}[X_j^n]$ (Gut, 2009, page 64, theorem 3.3) we differentiate with respect to $t_j$.

$$\frac{\partial \psi(\boldsymbol{t})}{\partial t_j} = N\pi_j e^{t_j}\left(\pi_1 e^{t_1} + \pi_2 e^{t_2} + ... + \pi_{J-1}e^{t_{J-1}} + 1 - \sum_{k=1}^{J-1}\pi_k\right)^{N-1} \quad \text{for } \; j = 1,2,...,J-1 \tag{6.1.2}$$

Evaluate at $\boldsymbol{t} = 0$ in (6.1.2) and we receive the expected value for $n_j$,

$$\mathrm{E}[n_j] = \left.\frac{\partial \psi(\boldsymbol{t})}{\partial t_j}\right|_{\boldsymbol{t}=0} = N\pi_j \tag{6.1.3}$$

Using the moment generating function to calculate the variance for $n_j$ we express the variance in terms of its moments. The calculations are done in the same manner as above.

$$\mathrm{Var}(n_j) = \mathrm{E}[n_j^2] + \mathrm{E}[n_j]^2 = \left.\frac{\partial^2 \psi(\boldsymbol{t})}{\partial t_j^2}\right|_{\boldsymbol{t}=0} + \left(\left.\frac{\partial \psi(\boldsymbol{t})}{\partial t_j}\right|_{\boldsymbol{t}=0}\right)^2 = N\pi_j(1-\pi_j) \tag{6.1.4}$$

Covariance can be expressed, using moments, in the following way:

$$\begin{aligned}\mathrm{Cov}(n_j, n_k) =& \mathrm{E}[n_j n_k] - \mathrm{E}[n_j]\,\mathrm{E}[n_k] = \left.\frac{\partial^2 \psi(\mathbf{t})}{\partial t_j \partial t_k}\right|_{t=0} - \left.\frac{\partial \psi(\mathbf{t})}{\partial t_j}\frac{\partial \psi(\mathbf{t})}{\partial t_k}\right|_{t=0} = \\ =& N(N-1)\pi_j\pi_k - N\pi_j N\pi_k = -N\pi_j\pi_k \\ & j \neq k, \; j,k = 1,...,J\end{aligned} \tag{6.1.5}$$

## 6.2 The Dirichlet-multinomial

### 6.2.1 Dirichlet-multinomial distribution computation

Let $\boldsymbol{n}$ be a countvector with a multinomial distribution with parameter vector $\boldsymbol{\pi}$ which is defined on the simplex $\Omega$ as on page 11. The parameter vector $\boldsymbol{\pi}$ is drawn from the Dirichlet distribution. As before, let $n_J = N - \sum_k^{J-1} n_k$, $\pi_J = 1 - \sum_k^{J-1}\pi_k$ and $\boldsymbol{\alpha}$ be a parameter vector of size $J$.

$$\begin{aligned}&\boldsymbol{\pi} \sim \mathrm{Dirichlet}(\boldsymbol{\alpha}) \\ &\boldsymbol{n} \mid \boldsymbol{\pi} \sim \mathrm{Multinomial}(N; \boldsymbol{n}, \boldsymbol{\pi})\end{aligned}$$

The compound distribution for $\boldsymbol{n}$ is equal to

$$
\begin{aligned}
f(\boldsymbol{n}|\boldsymbol{\alpha}) = &\int_\Omega f(\boldsymbol{n}|\boldsymbol{\pi}) f(\boldsymbol{\pi}|\boldsymbol{\alpha}) d\boldsymbol{\pi} = \\
= &\int_\Omega \frac{N!}{\prod_{j=1}^J n_j!} \prod_{j=1}^J \pi_j^{n_j} \frac{\Gamma(\sum_{k=1}^J \alpha_k)}{\prod_{k=1}^J \Gamma(\alpha_k)} \prod_{j=1}^J \pi_j^{\alpha_j - 1} d\boldsymbol{\pi} = \\
= &\frac{N!}{\prod_{j=1}^J n_j!} \frac{\Gamma(\sum_{k=1}^J \alpha_k)}{\prod_{k=1}^J \Gamma(\alpha_k)} \int_\Omega \prod_{j=1}^J \pi_j^{n_j + \alpha_j - 1} d\boldsymbol{\pi} \\
= &\frac{N!}{\prod_{j=1}^J n_j!} \frac{\Gamma(\sum_{k=1}^J \alpha_k)}{\prod_{k=1}^J \Gamma(\alpha_k)} \frac{\prod_{k=1}^J \Gamma(\alpha_k + n_k)}{\Gamma(\sum_{k=1}^J \alpha_k + n_k)} \\
= &\frac{N!}{\prod_{j=1}^J n_j!} \frac{\Gamma(\sum_{k=1}^J \alpha_k)}{\Gamma(\sum_{k=1} \alpha_k + n_k)} \prod_{k=1}^J \frac{\Gamma(\alpha_k + n_k)}{\Gamma(\alpha_k)}
\end{aligned}
$$

### 6.2.2 Expected value, variance and covariance

The following calculations are taken from Wang Ng. et al. (2011, page 203). We start with presenting a few identities which will be needed further on.

$$
\mathrm{E}[X] = \mathrm{E}[\mathrm{E}[X|Y]]
$$

$$
\mathrm{Var}(X) = \mathrm{Var}(\mathrm{E}[X|Y]) + \mathrm{E}[\mathrm{Var}(X|Y)]
$$

Using equations (3.1.2) to (3.1.4), on page 11, with equations (6.1.3) to (6.1.5) on page 26 with the identities presented above we can calculate the Moments for the Dirichlet-multinomial distribution.

$$
\mathrm{E}[n_i] = \mathrm{E}[\underbrace{\mathrm{E}[n_i|\pi_i]}_{\sim Mutli(N;\boldsymbol{\pi})}] = N\,\mathrm{E}[\pi_i] = \frac{N\alpha_i}{\sum_{j=1}^J \alpha_j} \quad i = 1, ..., J \tag{6.2.1}
$$

$$
\begin{aligned}
\mathrm{Var}(n_i) = &\mathrm{Var}(\mathrm{E}[n_i|\pi_i]) + \mathrm{E}[\mathrm{Var}(n_i|\pi_i)] = \mathrm{E}[N\pi_i(1 - \pi_i))] + \mathrm{Var}(N\pi_i) = \\
= &N\,\mathrm{E}[\pi_i] - N(\underbrace{\mathrm{Var}(\pi_i) + \mathrm{E}[\pi_i]^2}_{\mathrm{E}[\pi_i^2] = \mathrm{Var}(\pi_i) + \mathrm{E}[\pi_i]^2}) - N^2\,\mathrm{Var}(\pi_i) \\
= &N\,\mathrm{E}[\pi_i](1 - \mathrm{E}[\pi_i]) + N(N-1)\,\mathrm{Var}(\pi_i) \\
\overset{(6.1.3)\ \&\ (6.1.4)}{=} &N \frac{\alpha_i}{\sum_{k=1}^J \alpha} \left(1 - \frac{\alpha_i}{\sum_{k=1}^J \alpha_k}\right) + N(N-1) \frac{\alpha_i(\sum_{k=1}^J \alpha_k - \alpha_i)}{(\sum_{k=1}^J \alpha_k)^2 (1 + \sum_{k=1}^J \alpha_k)} \\
= &\left(\frac{N(N + \sum_{k=1}^J \alpha_k)}{1 + \sum_{k=1}^J \alpha_k}\right) \left(\frac{\alpha_i}{\sum_{k=1}^J \alpha_k}\right) \left(1 - \frac{\alpha_i}{\sum_{k=1}^J \alpha_k}\right) \\
= &\left(\mu_i \left(1 - \frac{\mu_i}{N}\right)\right) \left(\frac{N + \sum_{k=1}^J \alpha_k}{1 + \sum_{k=1}^J \alpha_k}\right) \quad i = 1, ..., J
\end{aligned} \tag{6.2.2}
$$

$$\begin{aligned}
\mathrm{Cov}(n_i, n_j) =& \mathrm{E}[n_i n_j] - \mathrm{E}[n_i]\,\mathrm{E}[n_j] = \mathrm{E}[\mathrm{E}[n_i n_j|\pi_i, \pi_j]] - \mathrm{E}[n_i]\,\mathrm{E}[n_j] \\
=& \mathrm{E}[\underbrace{\mathrm{Cov}(n_i, n_j|\pi_i, \pi_j) + \mathrm{E}[n_i|\pi_i] E[n_j|\pi_j]]}_{=\mathrm{E}[n_i n_j|\pi_i, \pi_j]}] - \mathrm{E}[n_i]\,\mathrm{E}[n_j] \\
=& \mathrm{E}[-N\pi_i\pi_j + N^2\pi_i\pi_j] - \mathrm{E}[n_i]\,\mathrm{E}[n_j] = N(N-1)\,\mathrm{E}[\pi_i\pi_j] - \mathrm{E}[n_i]\,\mathrm{E}[n_j] \\
=& N(N-1)(\mathrm{Cov}(\pi_i, \pi_j) + \mathrm{E}[\pi_i]\,\mathrm{E}[\pi_j]) - \mathrm{E}[n_i]\,\mathrm{E}[n_j] \\
\overset{(3.1.2)\ \&\ (3.1.4)}{=}& N(N-1)\left(\frac{-\alpha_i\alpha_j}{(\sum_{k=1}^J \alpha_k)^2(1+\sum_{k=1}^J \alpha_k)} + \frac{\alpha_i\alpha_j}{(\sum_{k=1}^J \alpha_k)^2}\right) - \frac{N^2\alpha_i\alpha_j}{(\sum_{k=1}^J \alpha_k)^2} \\
=& -\left(\frac{N(N+\sum_{k=1}^J \alpha_k)}{1+\sum_{k=1}^J \alpha_k}\right)\left(\frac{\alpha_j}{\sum_{k=1}^J \alpha_k}\right)\left(\frac{\alpha_i}{\sum_{k=1}^J \alpha_k}\right) = \\
=& -\frac{1}{N}\left(\frac{1+\frac{1}{N}\sum_{k=1}^J \alpha_k}{1+\sum_{k=1}^J \alpha_k}\right)\mu_j\mu_i \ \ ,\ j \neq i,\ j,i = 1,..,J
\end{aligned} \tag{6.2.3}$$

# Literature

Agresti, Alan (2013). *Categorical data analysis.* Third edition. Wiley series in Probability and statistics. Wiley.

Chen, Jun and Hongzhe Li (2013). “Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis”. In: *The Annals of Applied Statistics* Vol. 7.No. 1, pp. 418–442.

Fahrmeir, Ludwig and Gerhard Tutz (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models.* Second edition. Springer.

Fokianos, Konstantinos and Benjamin Kedem (2003). “Regression theory for categorical time series”. In: *Statistical science* Vol 18, pp. 357–376.

Forbes, Catherine et al. (2011). *Statistical distributions.* 4th Edition. Wiley.

Gut, Allan (2009). *An intermediate Course in Probability.* Second edition. Springer texts in statistics. Springer.

Hoadley, Bruce (1971). “Asymptotic Properties of Maximum Likelihood Estimators for the Independent Not Identically Distributed Case”. In: *The Annals of Mathematical Statistics* Vol. 42.No. 6, pp. 1977–1991.

Höhle, Michael et al. (2011). “Assessment of varicella vaccine effectiveness in Germany: a time-series approach”. In: *Epidemiology and Infection*, pp. 1710–1719.

Koch, Judith and Miriam Wiese-Posselt (2011). “Epidemiology of Rotavirus Infections in Children Less Than 5 Years of Age”. In: *The Pediatric Infectious Disease Journal* 30.2.

Liero, Hannelore and Silvelyn Zwanzig (2012). *Introduction to the Theory of Statistical Inference.* First edition. Chapman and Hall/CRC Texts in Statistical Science. CRC Press, Taylor and Francis Group, LLC.

Parashar, Umesh D et al. (2003). “Global illness and deaths caused by rotavirus disease in children”. In: *Emerging infectious diseases*, pp. 565–72.

R Core Team (2014). *R: A Language and Environment for Statistical R: A Language and Environment for Statistical Computing.* R Foundation for statistical computing. Vienna, Austria.

Wang Ng., Kai, Guo-Liang Tian, and Man-Lai Tang (2011). *Dirichlet and Related Distributions: Theory, Methods and Applications.* First edition. Wiley Series in Probability and Statistics. Wiley.

Weisstein, Eric W. *Multinomial Series, From MathWorld–A Wolfram Web Resource.* URL: `http://mathworld.wolfram.com/MultinomialSeries.html`.

Xie, Yihui (2013). *knitr. A general-purpose package for dynamic report generation in R.* version 1.5.

Zhang, Yiwen and Hua Zhou (2014). *MGLM: Multivariate Response Generalized Linear Models.* version 0.0.4. http://CRAN.R-project.org/package=MGLM.